Stochastic Processes and their Applications 48 (1993) 157–174
North-Holland

On residual sums of squares in non-parametric autoregression

B. Cheng and H. Tong

University of Kent at Canterbury, UK

Received 29 January 1992
Revised 24 September 1992

By relying on the theory of U-statistics of dependent data, we have given a detailed analysis of the residual sum of squares, RSS, after fitting a nonlinear autoregression using the kernel method. The asymptotic bias of the RSS as an estimator of the noise variance is evaluated up to and including the first order term. A similar quantity, the cross validated residual sum of squares obtained by 'leaving one out' in the fitting is similarly analysed. An asymptotic positive bias is obtained.

bias * cross validation * kernel * non-parametric autoregression * residual sum of squares * U-statistics

1. Introduction

Recently, a non-parametric approach based on the kernel method and others is making an increasing impact on nonlinear time series analysis. A recent reference is Auestad and Tjøstheim (1990), which has cited numerous important references. This development is quite natural in view of the rapid development in non-parametric regression with independent observations, the modern history of which probably goes back as far as Nadaraya (1965) and Watson (1964). Rosenblatt (1969) and Roussas (1969) give important early developments in the directions of regression and Markov sequence respectively. The book by Härdle (1990) is a useful reference.

Let us take as our non-linear autoregressive (NLAR) model

$$Z_t = F(Z_{t-1}, \ldots, Z_{t-d}) + \varepsilon_t, \tag{1.1}$$

where the autoregressive function F is *unknown* and $\{\varepsilon_t\}$ is a sequence of martingale difference with variance σ^2. Assume that $\{Z_t\}$ is a strictly stationary univariate time series with finite variance and absolutely continuous distribution. Let $(Z_1, \ldots, Z_N)$ denote the observations from (1.1).

To set up the kernel estimation, let $K_d : \mathbb{R}^d \to \mathbb{R}^1$ denote a non-negative even kernel on $\mathbb{R}^d$ which integrates to unity. Although not essential for our results, it is convenient to assume

$$K_d(u) = \prod_{i=1}^{d} k(u_i), \tag{1.2}$$

Correspondence to: Prof. H. Tong, Institute of Mathematics and Statistics, University of Kent, Cornwallis Building, Canterbury, Kent CT2 7NF, UK.

where $k:\mathbb{R}^1\to\mathbb{R}^1$ is a non-negative even kernel on $\mathbb{R}^1$ which integrates to unity and $u=(u_1,\ldots,u_d)\in\mathbb{R}^d$. Denote the row vector $(Z_{t-d},\ldots,Z_{t-1})$ by Y_t. Let f denote its density function. Let $r\ll N$ and $B(N)$ denote the bandwidth of the kernel. For $y\in\mathbb{R}^d$, define

$$\hat{f}_N(y)=\frac{1}{(N-r+1)(B(N))^d}\sum_{s=r}^{N}K_d\left(\frac{y-Y_s}{B(N)}\right), \tag{1.3}$$

$$\hat{F}_N(y)=\frac{1}{(N-r+1)(B(N))^d}\sum_{s=r}^{N}Z_sK_d\left(\frac{y-Y_s}{B(N)}\right)\{\hat{f}_N(y)\}^{-1}. \tag{1.4}$$

Let us define the (normalised) *residual sum of squares*, RSS, by

$$\mathrm{RSS}=\frac{1}{(N-r+1)}\sum_{t=r}^{N}\{Z_t-\hat{F}_N(Y_t)\}^2W(Y_t), \tag{1.5}$$

where W is a suitably chosen weight function and $r\geq d+1$. In the context of order determination, it would be necessary to replace RSS by RSS(d, $B(N)$) in preparation for a minimization with respect to d and $B(N)$. (See, e.g., Cheng and Tong (1992), 1993).) However, in our present context, d is fixed. To simplify our notation, we shall henceforth omit reference to d and N whenever there is no danger of confusion. If we envisage searching over d $(1\leq d\leq L)$, say, then we may set $r\geq L+1$. By analogy with classical regression theory, it is expected that RSS will have an asymptotic bias as an estimator of σ^2. Auestad and Tjøstheim (1990) have effectively conjectured that the relative bias is negative and equals $N^{-1}B^{-d}\gamma\int\{K(u)\}^2\,du$, where $\gamma=\int W(x)\,dx/\int W(x)f(x)\,dx$. To date, we know of no explicit evaluation of the bias to this order in the literature.

A measure related to the RSS is the *cross validated residual sum of squares*, CV:

$$\mathrm{CV}=\frac{1}{(N-r+1)}\sum_{t=r}^{N}\{Z_t-\hat{F}_{\setminus t}(Y_t)\}^2W(Y_t), \tag{1.6}$$

where $\hat{F}_{\setminus t}(y)$ and $\hat{f}_{\setminus t}(y)$ are as defined by (1.4) and (1.3) respectively, with the exception that now the summations over s omit t in each case and the divisor $(N-r+1)$ is replaced by $(N-r)$ for obvious reasons. (Note the omission of the suffix N—as announced earlier.) In the case of regression smoothing with independent observations, it is known that, for $d=1$, CV is a biased estimate of σ^2 with a relative positive bias of $2\{k(0)\}N^{-1}B^{-1}\gamma$. (See, e.g., Härdle, Hall and Marron (1988, equation 2.5).) Cheng and Tong (1992) have stated an extension of this result to nonlinear autoregression. We propose to give a complete proof of this result here. This result is significant because it provides the foundation for an order, i.e. d, determination of non-linear autoregression when the autoregressive function F is unknown. Cheng and Tong (op. cit) have given the details, and they have also proved the *consistency* of this CV and its related approaches (e.g. final prediction error-type estimates) under certain conditions.

2. Basic notations and assumptions

Let $\mathscr{A}_s^t(Z)$ denote $\sigma(Z_s,\ldots, Z_t)$, the sigma algebra generated by $(Z_s,\ldots, Z_t)$. Let $\mathscr{A}_s^t(\varepsilon) = \sigma(\varepsilon_s,\ldots, \varepsilon_t)$. Let $\{Z_t\}$ be a strictly stationary stochastic process with finite variance and absolutely continuous distribution. We assume that $\{Z_t\}$ is *absolutely regular*, i.e.

$$\beta_j = \sup_{i\in\mathbb{N}} E\left[\sup_{A\in\mathscr{A}_{i+j}^{\infty}(z)} \{|P(A|\mathscr{A}_1^i(Z)) - P(A)|\}\right] \to 0 \quad \text{as } j\to\infty. \tag{2.1}$$

The assumption of absolute regularity for model (1.1) is reasonable. For example, Pham and Tran (1985) and Mokkadem (1988) have considered absolute regularity for linear autoregressive/moving average models, which include the non-Gaussian cases. Pham (1986) has shown that under mild conditions bilinear models are absolutely regular, and Mokkadem (1987) has established the same for NLAR. In general, every strictly stationary real aperiodic Harris-recurrent Markov chain is absolutely regular (Bradley, 1986, p. 176). Note that uniform mixing implies absolute regularity, which implies strong (i.e. α) mixing (op. cit.)

For $\{\varepsilon_t\}$ we assume that $\forall t$,

(A1) $E[\varepsilon_t|\mathscr{A}_{-\infty}^{t-1}(Z)] = 0$ a.s., (2.2)

and

(A2) $E[\varepsilon_t^2|\mathscr{A}_{-\infty}^{t-1}(Z)] = \sigma^2$, a strictly positive constant, a.s. (2.3)

Assumption (A2) is slightly stronger than that required of martingale differences but is often used by Hannan and his associates in their studies of the problem of order determination in linear models. See, for example, Hannan and Kavalieris (1986) and An et al. (1982).

Now, we notice that from (1.1) and (2.2),

$$Z_t = E[Z_t|Z_{t-1},\ldots, Z_{t-d}] + \varepsilon_t, \tag{2.4}$$

with $E[\varepsilon_t|Z_{t-1},\ldots, Z_{t-d}] = 0$. That is, for each t, Z_t can be predicted optimally in the least squares sense in terms of $Z_{t-1},\ldots, Z_{t-d}$. Let $\tilde{Z}_t$ denote the least squares predictor given by the first term on the right-hand side of (2.4). Since F is not assumed known, we consider the Nadaraya-Watson kernel estimates such as $\hat{F}$ and $\hat{F}_{\setminus t}$ of the conditional expectation using observations $Z_1,\ldots, Z_N$.

(A3) $\int_{\mathbb{R}^1} |u|k(u)\,du < \infty$.

(A4) F is Hölder continuous, that is, $\forall x, y\in\mathbb{R}^d$,

$$|F(x) - F(y)| \leq c_1|x-y|^{\mu},$$

where $0<\mu\leq 1$ and $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^d$.

(A5) W is a weight funtion which has a compact support S, and

$$0<\int_{\mathbb{R}^d} W(x)\,dx<\infty, \quad 0\leqslant W(x)\leqslant 1.$$

(A6) f is strictly positive on S, and $\forall x,\ y\in\mathbb{R}^d$,

$$|f(x)-f(y)|\leqslant c_2|x-y|.$$

(A7) k has compact support, and $\forall x,\ y\in\mathbb{R}$,

$$|k(x)-k(y)|\leqslant c_3|x-y|.$$

We note that the compactness can be removed by using the method of function class as in Robinson (1988).

(A8) For every, t, s, τ, t', s', $\tau'\in\mathbb{N}$, the joint probability density function of $(Y_t, Y_s, Y_\tau, Y_{t'}, Y_{s'}, Y_{\tau'})$ is bounded. (We should mention that if, for example, $t=t'$, then the assumption requires that the joint probability density function of $(Y_t, Y_s, Y_\tau, Y_{s'}, Y_{\tau'})$ exists and is bounded.)

(A9) Let $p^{-1}+q^{-1}=1$. For some $p>2$ and $\delta>0$ such that $\delta<(2/q)-1$,

$$E|\varepsilon_1|^{2p(1+\delta)}<\infty \quad \text{and} \quad E|F(Y_1)|^{2p(1+\delta)}<\infty.$$

(A10) For δ in (A9),

$$\beta_j^{\delta/(1+\delta)}=O(j^{-2}),$$

where β_j is defined by (2.1).

(A11) Let $j=j(N)$ be a positive integer and $i=i(N)$ be the largest positive integer such that $2ij\leqslant N$ and

$$\limsup_{N\to\infty}(1+6e^{1/2}\beta_j^{1/(1+i)})^i<\infty.$$

(A12) For i in (A11) and the bandwidth B,

$$\limsup_{N\to\infty} iB^d<\infty.$$

(A13) $NB^{2d}\to\infty$ as $N\to\infty$.

(A14) For μ in (A4), $NB^{2d+2\mu}\to 0$ as $N\to\infty$.

Some explanation of the above conditions is in order. (A1)-(A4) are obvious. (A5) is the introduction of a weight function W, the purpose of which is to overcome the 'infinite integration problem' in asymptotic expansion encountered by Auestad and Tjøstheim (1990). (A6), (A7), (A9), (A13) and (A14) are standard conditions in non-parametric inference. (A8) is a mild condition, which will be useful when we use the mixing inequality. (A10) is a mild condition, which is weaker than geometric ergodicity. (See also comments prior to (A1).) (A11) and (A12) were given by Roussas (1988). They may be replaced by other assumptions on the mixing coefficient β, if other methods are used to show the almost sure consistency of $\hat{f}$ and $\hat{F}$.

3. The residual sum of squares

The following result will be useful.

Lemma 1 (Roussas, 1988, Theorem 3.1. *Under* (A3), (A6), (A7) *and* (A10)-(A13),

$$\sup_{x\in S}|\hat{f}(x)-f(x)| = o(1) \quad \text{a.s.} \qquad \square$$

Let $n = N-r+1$. Now, we state the first main result.

Theorem 1. *Under* (A1)-(A14),

$$\mathrm{RSS} = \sigma_N^2\{1-(2\alpha-\beta)\gamma\rho^d/N + o_p(\rho^d/N)\},$$

where

$$\sigma_N^2 = n^{-1}\sum_{t=r}^{N}\varepsilon_t^2 W(Y_t),$$

$$\alpha = K(0), \quad \beta = \int\{K(u)\}^2\,du, \quad \rho = \rho(N) = 1/B(N),$$

$$\gamma = \int W(x)\,dx \Big/ \int W(x)f(x)\,dx.$$

Proof. Following Härdle and Marron (1985), write

$$\hat{F}(Y_t)-F(Y_t) = (\hat{F}(Y_t)-F(Y_t))\hat{f}(Y_t)/f(Y_t) + (\hat{F}(Y_t)-F(Y_t))(f(Y_t)-\hat{f}(Y_t))/f(Y_t). \tag{3.1}$$

Note that by Lemma 1, the second term is negligible compared to the first. Since

$$\mathrm{RSS} = n^{-1}\sum_{t=r}^{N}\varepsilon_t^2 W(Y_t) + 2n^{-1}\sum_{t=r}^{N}\varepsilon_t(F(Y_t)-\hat{F}(Y_t))W(Y_t) + n^{-1}\sum_{t=r}^{N}(F(Y_t)-\hat{F}(Y_t))^2 W(Y_t),$$

by (3.1) and Lemma 1, we have

$$\begin{aligned}\mathrm{RSS} &= n^{-1}\sum_{t=r}^{N}\varepsilon_t^2 W(Y_t) + 2n^{-1}\sum_{t=r}^{N}\varepsilon_t(F(Y_t)-\hat{F}(Y_t))[\hat{f}(Y_t)/f(Y_t)]^2 W(Y_t) \\ &\quad + n^{-1}\sum_{t=r}^{N}(F(Y_t)-\hat{F}(Y_t))^2[\hat{f}(Y_t)/f(Y_t)]^2 W(Y_t) + o_p(R) \\ &= \mathrm{I}_{\mathrm{RSS}} + 2\mathrm{II}_{\mathrm{RSS}} + \mathrm{III}_{\mathrm{RSS}} + o_p(R), \quad \text{say,}\end{aligned} \tag{3.2}$$

where $R = 2\mathrm{II}_{\mathrm{RSS}} + \mathrm{III}_{\mathrm{RSS}}$. First, by an (i.e. any standard) ergodic theorem,

$$\mathrm{I}_{\mathrm{RSS}} = \sigma^2\int W(x)f(x)\,dx + o_p(1). \tag{3.3}$$

Now I_{RSS} is just σ_N^2. Secondly, we turn to II_{RSS}. Define

$$d_{t,s} = K(B^{-1}(Y_t - Y_s)), \tag{3.4}$$

$$c_{t,s} = (F(Y_t) - F(Y_s))d_{t,s}. \tag{3.5}$$

We get

$$\begin{aligned} II_{RSS} &= n^{-3}B^{-2d} \sum_{t,s,\tau=r}^{N} \varepsilon_t c_{t,s} d_{t,\tau} W(Y_t) f^{-2}(Y_t) \\ &\quad - n^{-3}B^{-2d} \sum_{t,s,\tau=r}^{N} \varepsilon_t \varepsilon_s d_{t,s} d_{t,\tau} W(Y_t) f^{-2}(Y_t) \\ &= II_{RSS}(1) - II_{RSS}(2), \quad \text{say.} \end{aligned}$$

Since

$$\mathscr{A}_{-\infty}^{t}(Z) \vee \mathscr{A}_{-\infty}^{t}(\varepsilon) = \mathscr{A}_{-\infty}^{t}(Z) \quad \text{and} \quad \mathscr{A}_{t+m}^{+\infty}(Z) \vee \mathscr{A}_{t+m}^{+\infty}(\varepsilon) \subset \mathscr{A}_{t+m-d}^{+\infty}(Z),$$

$e_t \triangleq (\varepsilon_t, Y_t)$, taking values in $\mathbb{R}^{d+1}$, is still absolutely regular. The decomposition (3.2) is in terms of sums of *dependent* observations. One way to analyse these is to symmetrize them and then appeal to the theory of U-statistics of dependent observations. To this end, define function $H^{(1)}$ by

$$\begin{aligned} H^{(1)}(e_t, e_s, e_\tau) &= \{\varepsilon_t d_{t,s} d_{t,\tau} (2F(Y_t) - F(Y_s) - F(Y_\tau)) W(Y_t) f^{-2}(Y_t) \\ &\quad + \varepsilon_s d_{s,t} d_{s,\tau} (2F(Y_s) - F(Y_t) - F(Y_\tau)) W(Y_s) f^{-2}(Y_s) \\ &\quad + \varepsilon_\tau d_{\tau,t} d_{\tau,s} (2F(Y_\tau) - F(Y_t) - F(Y_s)) W(Y_\tau) f^{-2}(Y_\tau)\}. \end{aligned}$$

Then

$$II_{RSS}(1) = \tfrac{1}{6} n^{-3} B^{-2d} \sum_{t,s,\tau=r}^{N} H^{(1)}(e_t, e_s, e_\tau), \tag{3.6}$$

where the function $H^{(1)}$ is symmetric in its three arguments.

Let G, $G_{t,s}$ and $G_{t,s,\tau}$ be the distributions of e_t, (e_t, e_s) and (e_t, e_s, e_τ) respectively. Define

$$H_1^{(1)}(x_1) = \int H^{(1)}(x_1, x_2, x_3)\, dG(x_2) \times dG(x_3),$$

$$H_2^{(1)}(x_1, x_2) = \int H^{(1)}(x_1, x_2, x_3)\, dG(x_3).$$

Mimicking Hoeffding's projection method (see, e.g. Denker and Keller, 1983), we have

$$\begin{aligned} II_{RSS}(1) &= B^{-2d}\tilde{R} + \tfrac{1}{2} n^{-1} B^{-2d} \sum_{t=r}^{N} H_1^{(1)}(e_t) \\ &\quad + \tfrac{1}{2} n^{-3} B^{-2d} \sum_{t,s=r}^{N} [H^{(1)}(e_t, e_s, e_s) - 3H_2^{(1)}(e_t, e_s) + 3H_1^{(1)}(e_t)] \\ &\quad + \tfrac{1}{2} n^{-2} B^{-2d} \sum_{t=r}^{N} [H_2^{(1)}(e_t, e_t) - 2H_1^{(1)}(e_t)], \quad \text{say.} \end{aligned} \tag{3.7}$$

The decomposition os $\mathrm{II}_{\mathrm{RSS}}(1)$ in (3.7) is motivated by the desire to show that $\mathrm{II}_{\mathrm{RSS}}(1)$ is negligible. This is achieved by showing that each of the terms on the right-hand side of (3.7) is negligible. Let $\{e_t^{(1)}\}$, $\{e_t^{(2)}\}$ and $\{e_t^{(3)}\}$ be multiple independent copies of $\{e_t\}$. For convenience, let $e_t^{(0)}$ denote e_t. Then, by Denker and Keller's Proposition 2 (1983), the residual term $\tilde{R}$ in (3.7) satisfies

$$E\tilde{R}^2 \leq cN^{-2}\{\sup E\left|H^{(1)}(e_t^{(i_1)}, e_s^{(i_2)}, e_\tau^{(i_3)})H^{(1)}(e_{t'}^{(j_1)}, e_{s'}^{(j_2)}, e_{\tau'}^{(j_3)})\right|^{1+\delta}\}^{1/(1+\delta)}$$
$$+ cN^{-3}\sup\{E\left|H^{(1)}(e_t^{(i_1)}, e_s^{(i_2)}, e_\tau^{(i_3)})H^{(1)}(e_{t'}^{(j_1)}, e_{s'}^{(j_2)}, e_{\tau'}^{(j_3)})\right|^{1+\delta}\}^{1/(1+\delta)},$$

where the first supremum is over all (i_1, i_2, i_3) and $(j_1, j_2, j_3) \in \{0, 1, 2, 3\}^3$ and all $t, s, \tau, t', s', \tau'$ such that $t \neq s$, $t \neq \tau$, $\tau \neq s$, $t' \neq s'$, $t' \neq \tau'$, $\tau' \neq s'$ and, at most, only one pair of co-ordinates between (t, s, τ) and (t', s', τ') is equal, and the second supremum is over all (i_1, i_2, i_3) and $(j_1, j_2, j_3) \in \{0, 1, 2, 3\}^3$ and all $t, s, \tau, t', s', \tau'$ such that $t \neq s$, $t \neq \tau$, $\tau \neq s$, $t' \neq s'$, $t' \neq \tau'$, $\tau' \neq s'$ and, at least, two pairs of co-ordinates between (t, s, τ) and (t', s', τ') are equal.

By Hölder's inequality, (A8), (A9), the first supremum is bounded by

$$B^{4d/(q+q\delta)}\{E\left|\varepsilon_t\right|^{2p(1+\delta)}\}^{1/(2p)}\{E\left|F(Y_t)\right|^{2p(1+\delta)}\}^{1/(2p)} \tag{3.8}$$

for $p^{-1}+q^{-1}=1$ and $q(1+\delta)<2$, but p need not be greater than q, and the second supremum is bounded by

$$B^{2d/(q+q\delta)}\{E\left|\varepsilon_t\right|^{2p(1+\delta)}\}^{p/2}\{E\left|F(Y_t)\right|^{2p(1+\delta)}\}^{p/2}. \tag{3.9}$$

Thus, by (3.8), (3.9), and (A13),

$$\tilde{R} = o_p((NB^d)^{-1}). \tag{3.10}$$

Next, we treat the second term in (3.7). Let

$$g_t = \int d_{t,s}d_{t,\tau}(F(Y_t) - F(Y_s))\,\mathrm{d}G \times \mathrm{d}G.$$

We have

$$H_1^{(1)}(e_t) = 2\varepsilon_t g_t W(Y_t) f^{-2}(Y_t).$$

Hence,

$$E[n^{-1}B^{-2d}\sum_{t=r}^{N} H_1^{(1)}(e_t)]^2 \leq c\sigma^2 n^{-1}B^{-4d}E[g_t^2 W^2(Y_t)].$$

Since F is Hölder continuous, by (A3), (A4) and the boundedness and compactness of K,

$$|g_t| \leq cB^{\mu+2d}.$$

We get by (A14),

$$n^{-1}B^{-2d}\sum_{t=r}^{N} H_1^{(1)}(e_t) = o_p((NB^d)^{-1}). \tag{3.11}$$

(*Note*: As a referee has pointed out, the right side of (3.11) may be sharpened to $o_p((NB^{2d+\mu})^{-1})$. However, the cruder term is sufficient for Theorem 1.) Now, we treat the third term in (3.7). We know that, while $s = \tau$,

$$H^{(1)}(e_t, e_s, e_s) = 2\varepsilon_t d_{t,s}^2 (F(Y_t) - F(Y_s)) W(Y_t) f^{-2}(Y_t)) + 2K(0)\varepsilon_s d_{t,s}(F(Y_s) - F(Y_t)) W(Y_s) f^{-2}(Y_s).$$

By $E[\varepsilon_t | \mathscr{A}_{-\infty}^{t-1}(Z)] = E[\varepsilon_s | \mathscr{A}_{\infty}^{s-1}(Z)] = 0$, (A3), (A4), (A6), (A7), it is easy to see

$$E\left\{\sum_{t,s=r}^{N} [H^{(1)}(e_t, e_s, e_s) - 3H_2^{(1)}(e_t, e_s) + 3H_1^{(1)}(e_t)]\right\}^2 \leqslant cN^3.$$

So by (A13),

$$\tfrac{1}{2}n^{-3}B^{-2d} \sum_{t,s=r}^{N} [H^{(1)}(e_t, e_s, e_s) - 3H_2^{(1)}(e_t, e_s) + 3H_1^{(1)}(e_t)] = o_p((NB^d)^{-1}). \tag{3.12}$$

Similarly, by

$$H_2^{(1)}(e_t, e_t) = 2\varepsilon_t K(0) \int d_{t,\tau}(F(Y_t) - F(Y_\tau))\, \mathrm{d}GW(Y_t) f^{-2}(Y_t)$$

and

$$H_1^{(1)}(e_t) = 2\varepsilon_t \int d_{t,s} d_{t,\tau} (2F(Y_t) - F(Y_s) - F(Y_\tau))\, \mathrm{d}G \times \mathrm{d}GW(Y_t) f^{-2}(Y_t),$$

from $E[\varepsilon_t | \mathscr{A}_{-\infty}^{t-1}] = 0$ and (A3), (A4), (A6), (A7),

$$E\left\{\sum_{t=r}^{N} [H_2^{(1)}(e_t, e_t) - 2H_1^{(1)}(e_t)]\right\}^2 \leqslant cN.$$

So by (A13), we have

$$\tfrac{1}{2}n^{-2}B^{-2d} \sum_{t=r}^{N} [H_2^{(1)}(e_t, e_t) - 2H_1^{(1)}(e_t)] = o_p((NB^d)^{-1}). \tag{3.13}$$

Therefore, putting everything together, we get

$$\mathrm{II}_{\mathrm{RSS}}(1) = o_p((NB^d)^{-1}). \tag{3.14}$$

Define

$$H^{(2)}(e_t, e_s, e_\tau) = (\varepsilon_t\varepsilon_s + \varepsilon_t\varepsilon_\tau) d_{t,s} d_{t,\tau} W(Y_t) f^{-2}(Y_t) + (\varepsilon_s\varepsilon_t + \varepsilon_s\varepsilon_\tau) d_{s,t} d_{s,\tau} W(Y_s) f^{-2}(Y_s) + (\varepsilon_\tau\varepsilon_s + \varepsilon_\tau\varepsilon_t) d_{\tau,s} d_{\tau,t} W(Y_\tau) f^{-2}(Y_\tau).$$

Similar to $H_i^{(1)}$, $i=1,2$, we may define $H_i^{(2)}$, $i=1,2$. Hence

$$\begin{aligned}\mathrm{II}_{\mathrm{RSS}}(2) &= \tfrac{1}{6}n^{-3}B^{-2d}\sum_{t,s,\tau=r}^{N} H^{(2)}(e_t, e_s, e_\tau)\\ &= \tfrac{1}{6}n^{-3}B^{-2d}\sum_{t,s,\tau=r}^{N}[H^{(2)}(e_t, e_s, e_\tau) - 3H_2^{(2)}(e_t, e_s)]\\ &\quad + \tfrac{1}{2}n^{-2}B^{-2d}\sum_{t,s,t=r}^{N} H_2^{(2)}(e_t, e_s).\end{aligned}$$

Notice here $H_1^{(2)}(e_t)\equiv 0$. By similar treatment for $\mathrm{II}_{\mathrm{RSS}}(1)$, we can show that

$$\tfrac{1}{6}n^{-3}B^{-2d}\sum_{t,s,\tau=r}^{N}[H^{(2)}(e_t, e_s, e_\tau) - 3H_2^{(2)}(e_t, e_s)] = \mathrm{o_p}((NB^d)^{-1}).$$

On the other hand,

$$\begin{aligned}H_2^{(2)}(e_t, e_s) &= \varepsilon_t\varepsilon_s\int d_{s,t}d_{s,\tau}\,\mathrm{d}GW(Y_s)f^{-2}(Y_s)\\ &\quad + \varepsilon_t\varepsilon_s\int d_{t,s}d_{t,\tau}\,\mathrm{d}GW(Y_t)f^{-2}(Y_t).\end{aligned}$$

Hence

$$\begin{aligned}&\tfrac{1}{2}n^{-2}B^{-2d}\sum_{t,s=r}^{N} H_2^{(2)}(e_t, e_s)\\ &= n^{-2}B^{-2d}\sum_{t=r}^{N}\varepsilon_t^2\int d_{t,t}d_{t,\tau}\,\mathrm{d}GW(Y_t)f^{-2}(Y_t)\\ &\quad + 2n^{-2}B^{-2d}\sum_{s<t}\varepsilon_s\varepsilon_t\int d_{s,t}d_{s,\tau}\,\mathrm{d}GW(Y_s)f^{-2}(Y_s)\\ &= \mathrm{II}_{\mathrm{RSS}}(2,1) + 2\mathrm{II}_{\mathrm{RSS}}(2,2), \quad \text{say.}\end{aligned}$$

Firstly,

$$\begin{aligned}\mathrm{II}_{\mathrm{RSS}}(2,1) &= n^{-2}B^{-d}K(0)\sum_{t=r}^{N}\varepsilon_t^2 W(Y_t)f^{-1}(Y_t)\\ &\quad + K(0)B^d\sum_{t=r}^{N}\varepsilon_t^2[B^{-d}\int d_{t,\tau}\,\mathrm{d}G - f(Y_t)]W(Y_t)f^{-2}(Y_t).\end{aligned}$$

Since f is Hölder continuous, we have

$$E\left|B^{-d}\int d_{t,\tau}\mathrm{d}G - f(Y_t)\right| = \mathrm{O}(B).$$

So, we have

$$\mathrm{II}_{\mathrm{RSS}}(2,1) = (NB^d)^{-1}K(0)\sigma_N^2\gamma_N + \mathrm{o_p}((NB^d)^{-1}),$$

where

$$\gamma_N = n^{-1} \sum_{t=r}^{N} \varepsilon_t^2 W(Y_t) f^{-1}(Y_t)/\sigma_N^2.$$

By an (i.e. any standard) ergodic theorem, we have

$$\gamma_N = \int W(x)\,dx \Big/ \int W(x) f(x)\,dx + o_p(1)$$
$$= \gamma + o_p(1), \quad \text{say.}$$

Hence, we get

$$\mathrm{II}_{\mathrm{RSS}}(2,1) = (NB^d)^{-1} K(0)\gamma\sigma_N^2 + o_p((NB^d)^{-1}). \tag{3.15}$$

Secondly, we define

$$b_{s,t} = b(Y_s, Y_t) \triangleq \int d_{s,t} d_{s,\tau}\,dG W(Y_s) f^{-2}(Y_s).$$

So

$$E\left[\sum_{s<t} \varepsilon_s \varepsilon_t \int d_{s,t} d_{s,\tau}\,dG W(Y_s) f^{-2}(Y_s)\right]^2$$
$$= \sigma^2 \sum{}' E\left[\varepsilon_{s'} \varepsilon_s b_{s',t} b_{s,t} - \varepsilon_{s'} \varepsilon_s \int b_{s',t} b_{s,t}\,dG\right] + \sigma^2 \sum_{s<t} E[\varepsilon_s^2 b_{s,t}^2],$$

where $\sum'$ denotes a summation over $s<t$, $s'<t$ and $s \neq s'$. By Denker and Keller's Lemma 6 (1983), for $s' \neq s$, we have

$$\left|E[\varepsilon_{s'} \varepsilon_s b_{s',t} b_{s,t} - \varepsilon_{s'} \varepsilon_s \int b_{s',t} b_{s,t}\,dG]\right|$$
$$\leq 4\beta_{t-\max\{s,s'\}}^{\delta/(1+\delta)} \max\{E|\varepsilon_{s'} \varepsilon_s b_{s',t} b_{s,t}|^{1+\delta},$$
$$E|\varepsilon_{s'} \varepsilon_s b(Y_s, Y_t^{(1)}) b(Y_s, Y_t^{(1)})|^{1+\delta}\}^{1/(1+\delta)}$$
$$\leq c\beta_{t-\max\{s,s'\}}^{\delta/(1+\delta)} B^{4d/q(1+\delta)} \quad \text{(by (A8))}.$$

Hence by (A10),

$$\sum{}' \left|E[\varepsilon_{s'} \varepsilon_s b_{s',t} b_{s,t} - \varepsilon_{s'} \varepsilon_s \int b_{s',t} b_{s,t}\,dG]\right| = O(NB^{4d/(q(1+\delta))}).$$

And by (A2), (A8),

$$\sum_{s<t} E[\varepsilon_s^2 b_{s,t}^2] = \sum_{s<t} \sigma^2 E b_{s,t}^2 = O(N^2 B^{3d}).$$

So, we get

$$\mathrm{II}_{\mathrm{RSS}}(2,2) = o_p((NB^d)^{-1}). \tag{3.16}$$

Finally, we consider the term $\mathrm{III}_{\mathrm{RSS}}$. By (3.2),

$$\begin{aligned}\mathrm{III}_{\mathrm{RSS}} &= n^{-3}B^{-2d}\sum_{t,s,\tau=r}^{N}\varepsilon_s\varepsilon_\tau d_{t,s}d_{t,\tau}W(Y_t)f^{-2}(Y_t)\\ &\quad -2n^{-3}B^{-2d}\sum_{t,s,\tau=r}^{N}\varepsilon_s d_{t,s}c_{t,\tau}W(Y_t)f^{-2}(Y_t)\\ &\quad +n^{-3}B^{-2d}\sum_{t,s,\tau=r}^{N}c_{t,s}c_{t,\tau}W(Y_t)f^{-2}(Y_t)\\ &= \mathrm{III}_{\mathrm{RSS}}(1)-2\mathrm{III}_{\mathrm{RSS}}(2)+\mathrm{III}_{\mathrm{RSS}}(3), \quad \text{say.}\end{aligned}$$

Similar to $\mathrm{II}_{\mathrm{RSS}}(2,1)$, we can show that

$$\mathrm{III}_{\mathrm{RSS}}(1)=(NB^d)^{-1}\int\{K(u)\}^2\,\mathrm{d}u\,\sigma_N^2\gamma+\mathrm{o_p}((NB^d)^{-1}). \tag{3.17}$$

Using similar treatment for $\mathrm{II}_{\mathrm{RSS}}(1)$, we get

$$\mathrm{III}_{\mathrm{RSS}}(2)=\mathrm{o_p}((NB^d)^{-1}. \tag{3.18}$$

Finally, by (A4),

$$E[|\mathrm{III}_{\mathrm{RSS}}(3)|]\leq cB^{2\mu}. \tag{3.19}$$

So, by (A14),

$$\mathrm{III}_{\mathrm{RSS}}(3)=\mathrm{o_p}((NB^d)^{-1}). \qquad \square$$

Suppose that we choose the kernel to be of the Gaussian type,

$$k(u)=(2\pi)^{-1/2}\exp(-0.5u)^2. \tag{3.20}$$

Then

$$\alpha=(2\pi)^{-d/2}, \qquad \beta=(4\pi)^{-d/2}.$$

For this choice of k the bias as quantified by Theorem 1 is indeed negative as conjectured by Auestad and Tjøstheim (1990). In a recent unpublished preprint, they have revised the magnitude of the negative bias, which for our set-up is now in agreement with ours. The said preprint has also stated a result based on heuristics which covers the case of non-constant conditional variance, i.e. relaxes (A2).

4. Cross validation

The second main result is stated in the form of the following theorem.

Theorem 2. *Under* (A1)–(A14),

$$\mathrm{CV}=\mathrm{RSS}\{1+2\alpha\gamma\rho^d/N+\mathrm{o_p}(\rho^d/N)\}.$$

Remark. A similar result was obtained by Härdle et al. (1988) in the nonparametric regression model.

Proof of Theorem 2. Let all summations be from r to N.

$$\begin{aligned}\mathrm{CV} &= n^{-1}\sum_t [Z_t - \hat{F}(Y_t) + \hat{F}(Y_t) - \hat{F}_{\setminus t}(Y_t)]^2 W(Y_t)\\ &= \mathrm{RSS} + 2n^{-1}\sum_t [Z_t - \hat{F}(Y_t)][\hat{F}(Y_t) - \hat{F}_{\setminus t}(Y_t)] W(Y_t)\\ &\quad + n^{-1}\sum_t [\hat{F}(Y_t) - \hat{F}_{\setminus t}(Y_t)]^2 W(Y_t)\\ &= \mathrm{RSS} + 2\mathrm{I}^*_{\mathrm{CV}} + \mathrm{II}^*_{\mathrm{CV}}, \quad \text{say.}\end{aligned} \tag{4.1}$$

Since

$$\hat{f}(Y_t) - \hat{f}_{\setminus i}(Y_t) = -\frac{1}{n-1}\hat{f}_{\setminus i}(Y_t) + \frac{1}{(n-1)B^d} K\left(\frac{Y_t - Y_i}{B}\right),$$

by the boundedness of K and f and Lemma 1,

$$\hat{f}(Y_t) - \hat{f}_{\setminus i}(Y_t) \sim (NB^d)^{-1}. \tag{4.2}$$

Since $\mathrm{I}^*_{\mathrm{CV}}$ contains random denominators, we shall employ the usual technique in kernel estimation (e.g. Härdle and Marron, 1985) to remove them. This, together with (4.2) and Lemma 1, gives

$$\begin{aligned}\mathrm{I}^*_{\mathrm{CV}} &= n^{-1}\sum_t [Z_t\hat{f}(Y_t) - \hat{F}(Y_t)\hat{f}(Y_t)]\\ &\qquad [\hat{F}(Y_t)\hat{f}(Y_t) - \hat{F}_{\setminus t}(Y_t)\hat{f}_{\setminus t}(Y_t)] f^{-2}(Y_t) W(Y_t) + \mathrm{o_p}(I^*_{\mathrm{CV}})\\ &= \mathrm{I}_{\mathrm{CV}} + \mathrm{o_p}(\mathrm{I}_{\mathrm{CV}}), \quad \text{say,}\end{aligned}$$

and

$$\begin{aligned}\mathrm{II}^*_{\mathrm{CV}} &= n^{-1}\sum_t [\hat{F}(Y_t)\hat{f}(Y_t) - \hat{F}_{\setminus t}(Y_t)\hat{f}_{\setminus t}(Y_t)]^2 f^{-2}(Y_t) W(Y_t) + \mathrm{o_p}(\mathrm{II}^*_{\mathrm{CV}})\\ &= \mathrm{II}_{\mathrm{CV}} + \mathrm{o_p}(\mathrm{II}_{\mathrm{CV}}), \quad \text{say.}\end{aligned}$$

Using

$$\begin{aligned}\hat{F}(Y_t)\hat{f}(Y_t) - \hat{F}_{\setminus t}(Y_t)\hat{f}_{\setminus t}(Y_t) &= n^{-1}B^{-d}\sum_s Z_s d_{t,s} - (n-1)^{-1}B^{-d}\sum_{s\neq t} Z_s d_{t,s}\\ &= -n^{-1}\hat{F}_{\setminus t}(Y_t)\hat{f}_{\setminus t}(Y_t) + n^{-1}B^{-d}Z_t d_{t,t},\end{aligned} \tag{4.3}$$

and

$$\begin{aligned}Z_t\hat{f}(Y_t) - \hat{F}(Y_t)\hat{f}(Y_t) &= n^{-1}B^{-d}\sum_s (Z_t - Z_s) d_{t,s}\\ &= n^{-1}B^{-d}\sum_s (\varepsilon_t - \varepsilon_s) d_{t,s} + n^{-1}B^{-d}\sum_s c_{t,s},\end{aligned} \tag{4.4}$$

we get

$$\begin{aligned} I_{CV} &= \alpha n^{-3} B^{-2d} \sum_{t,s} (\varepsilon_t - \varepsilon_s) d_{t,s} Z_t f^{-2}(Y_t) W(Y_t) \\ &\quad + \alpha n^{-3} B^{-2d} \sum_{t,s} Z_t c_{t,s} f^{-2}(Y_t) W(Y_t) \\ &\quad - n^{-3} B^{-d} \sum_{t,s} (\varepsilon_t - \varepsilon_s) d_{t,s} \hat{F}_{\setminus t}(Y_t) \hat{f}_{\setminus t}(Y_t) f^{-2}(Y_t) W(Y_t) \\ &\quad - n^{-3} B^{-d} \sum_{t,s} c_{t,s} \hat{F}_{\setminus t}(Y_t) \hat{f}_{\setminus t}(Y_t) f^{-2}(Y_t) W(Y_t) \\ &= I_{CV}(1) + I_{CV}(2) - I_{CV}(3) - I_{CV}(4), \quad \text{say.} \end{aligned} \tag{4.5}$$

$$\begin{aligned} I_{CV}(1) &= \alpha n^{-3} B^{-2d} \sum_{t,s} (\varepsilon_t - \varepsilon_s) d_{t,s} \varepsilon_t f^{-2}(Y_t) W(Y_t) \\ &\quad + \alpha n^{-3} B^{-2d} \sum_{t,s} (\varepsilon_t - \varepsilon_s) d_{t,s} F(Y_t) f^{-2}(Y_t) W(Y_t) \\ &= I_{CV}(1,1) + I_{CV}(1,2), \quad \text{say.} \end{aligned} \tag{4.6}$$

Let

$$\begin{aligned} H^{(1)}(e_t, e_s) &= (\varepsilon_t - \varepsilon_s) d_{t,s} \varepsilon_t f^{-2}(Y_t) W(Y_t) \\ &\quad + (\varepsilon_s - \varepsilon_t) d_{s,t} \varepsilon_s f^{-2}(Y_s) W(T_s), \end{aligned} \tag{4.7}$$

$$H_1^{(1)}(x_1) = \int H^{(1)}(x_1, x_2)\, \mathrm{d}G(x_2), \tag{4.8}$$

and

$$H_0^{(1)} = \int H^{(1)}(x_1, x_2)\, \mathrm{d}G(x_1) \times \mathrm{d}G(x_2). \tag{4.9}$$

Then

$$H_1^{(1)}(e_t) = \varepsilon_t^2 f^{-2}(Y_t) W(Y_t) \int d_{t,s}\, \mathrm{d}G + \sigma^2 \int d_{s,t} f^{-2}(Y_t) W(Y_s)\, \mathrm{d}G \tag{4.10}$$

and

$$H_0^{(1)} = 2\sigma^2 \int d_{t,s} f^{-2}(Y_t) W(Y_t)\, \mathrm{d}G \times \mathrm{d}G. \tag{4.11}$$

Similar to (3.7), we have

$$\begin{aligned} I_{CV}(1,1) &= \tfrac{1}{2} \alpha n^{-3} B^{-2d} \sum_{t,s} [H^{(1)}(e_t, e_s) - 2H_1^{(1)}(e_t) + H_0^{(1)}] \\ &\quad + \alpha n^{-2} B^{-2d} \sum_{t} [H_1^{(1)}(e_t) - H_0^{(1)}] + \tfrac{1}{2} \alpha n^{-1} B^{-2d} H_0^{(1)}. \end{aligned} \tag{4.12}$$

By a similar treatment to $II_{RSS}(1)$, we can show that the first two terms of (4.12)

are each $o_p((NB^d)^{-1})$. On the other hand,

$$\int d_{t,s} f^{-2}(Y_t) W(Y_t)\, dG \times dG$$

$$= \int K(B^{-1}(x_1 - x_2)) f^{-1}(x_1) f(x_2) W(x_1)\, dx_1\, dx_2$$

$$= B^d \int W(x)\, dx + o_p(B^d) \quad \text{(by (1.2))}. \tag{4.13}$$

So

$$I_{CV}(1, 1) = \alpha n^{-1} B^{-d} \int W(x)\, dx + o_p((NB^d)^{-1}). \tag{4.14}$$

By a similar treatment to $II_{RSS}(1)$, we can show that

$$I_{CV}(1, 2) = o_p((NB^d)^{-1}). \tag{4.15}$$

Let

$$H^{(2)}(e_t, e_s) = Z_t c_{t,s} f^{-2}(Y_t) W(Y_t) + Z_s c_{s,t} f^{-2}(Y_s) W(Y_s),$$

$$H_1^{(2)}(x_1) = \int H^{(2)}(x_1, x_2)\, dG(x_2),$$

$$H_0^{(2)} = \int H^{(2)}(x_1, x_2)\, dG(x_1)\, dG(x_2).$$

Then we can write

$$I_{CV}(2) = \tfrac{1}{2}\alpha n^{-3} B^{-2d} \sum_{t,s} H^{(2)}(e_t, e_s)$$

$$= \tfrac{1}{2}\alpha n^{-3} B^{-2d} \sum_{t,s} [H^{(2)}(e_t, e_s) - 2H_1^{(2)}(e_t) + H_0^{(2)}]$$

$$+ \alpha n^{-2} B^{-2d} \sum_{t} [H_1^{(2)}(e_t) - H_1^{(2)}] + \tfrac{1}{2}\alpha n^{-1} B^{-2d} H_0^{(2)}$$

$$= I_{CV}(2, 1) + \tfrac{1}{2}\alpha n^{-1} B^{-2d} H_0^{(2)}, \quad \text{say.} \tag{4.16}$$

By a similar treatment to $II_{RSS}(1)$, we can show that

$$I_{CV}(2, 1) = o_p((NB)^{-1}). \tag{4.17}$$

On the other hand, by (A1),

$$H_0^{(2)} = 2 \int F(x_1)[F(x_1) - F(x_2)] K(B^{-1}(x_1 - x_2)) f^{-1}(x_1) f(x_2) W(x_1)\, dx_1\, dx_2$$

$$\leqslant \text{const} \int |F(x_1)| |x_1 - x_2|^{\mu} K(B^{-1}(x_1 - x_2)) f^{-1}(x_1) f(x_2) W(x_1)\, dx_1\, dx_2$$

(by (A4))

$$\sim B^{\mu+d} \int |F(x_1)| |x_3|^{\mu} K(x_3) f^{-1}(x_1) f(x_1 - Bx_3) W(x_1)\, dx_1\, dx_3$$

$$\sim B^{\mu+d} \quad \text{as } N \to \infty \quad \text{(by (A5) and (A6))}. \tag{4.18}$$

Therefore, we have

$$\tfrac{1}{2}\alpha n^{-1}B^{-2d}H_0^{(2)} = o_p((NB^d)^{-1}, \tag{4.19}$$

and hence

$$I_{CV}(2) = o_p((NB^d)^{-1}). \tag{4.20}$$

Now

$$\begin{aligned} I_{CV}(3) &= n^{-3}(n-1)^{-1}B^{-2d}\sum_{\substack{t,s,\tau=r\\ t\neq\tau}}^{N}(\varepsilon_t-\varepsilon_s)\varepsilon_\tau d_{t,s}d_{t,\tau}f^{-2}(Y_t)W(Y_t) \\ &\quad + n^{-3}(n-1)^{-1}B^{-2d}\sum_{\substack{t,s,\tau=r\\ t\neq\tau}}^{N}(\varepsilon_t-\varepsilon_s)F(Y_\tau)d_{t,s}d_{t,\tau}f^{-2}(Y_t)W(Y_t) \\ &= n^{-1}[I_{CV}(3,1)+I_{CV}(3,2)], \quad \text{say.} \end{aligned} \tag{4.21}$$

Using a similar argument as previously, we have

$$I_{CV}(3,1) = o_p((NB^d)^{-1}) \quad \text{and} \quad I_{CV}(3,2) = o_p((NB)^{-1}).$$

Thus

$$I_{CV}(3) = o_p((NB^d)^{-1}).$$

Similarly

$$I_{CV}(4) = o_p((NB^d)^{-1}).$$

Putting everything together, we have proved that

$$I_{CV} = \alpha\sigma^2 n^{-1}\rho^d\int W(x)\,dx + o_p((NB^d)^{-1}). \tag{4.22}$$

By Theorem 1, RSS is a consistent estimator of $\sigma^2\int W(x)f(x)\,dx$. Therefore

$$I_{CV} = \mathrm{RSS}\{\alpha\gamma\rho^d/N + o_p(\rho^d/N)\}.$$

Next,

$$\begin{aligned} II_{CV} &= n^{-1}\sum_t[n^{-1}B^{-d}Z_t d_{t,t} - n^{-1}\hat{F}_{\setminus t}(Y_t)\hat{f}_{\setminus t}(Y_t)]^2 f^{-2}(Y_t)W(Y_t) \\ &\leq 2n^{-3}B^{-2d}\sum_t Z_t^2 d_{t,t}^2 f^{-2}(Y_t)W(Y_t) \\ &\quad + 2n^{-3}\sum_t[\hat{F}_{\setminus t}(Y_t)\hat{f}_{\setminus t}(Y_t)]^2 f^{-2}(Y_t)W(Y_t). \end{aligned} \tag{4.23}$$

It easily follows that

$$II_{CV} = o_p((NB^d)^{-1}). \tag{4.24}$$

Equations (4.1), (4.22) and (4.24) together then complete the proof of the theorem. □

5. Discussion

There is a point of contact between our results and those of Gjörfi, Härdle, Sarda and Vieu (1990, especially Chapter 6), namely the same asymptotic expression for the rate of convergence, i.e. $(NB^d)^{-1}$, and both groups use the cross-validation approach. However, the motivations are different. Their analysis is mainly motivated by the optimal choice of the bandwidth, i.e. B, whilst ours is the consistent determination of the number of 'regressors', i.e. d. For the latter problem, an exact asymptotic expression such as that given by Theorem 2 is critical. Otherwise the proof could be much simplified and the assumptions weakened. Note that the assumption of absolute regularity is not crucial for our results. For example, it may be replaced by strong mixing. (See, e.g., Cheng and Tong, 1993.)

Theorems 1 and 2 imply that

$$\mathrm{CV} = \sigma_N^2\{1 + \beta\gamma\rho^d / N + \mathrm{o_p}(\rho^d / N)\},$$

and so

$$E[\mathrm{CV}] = \sigma^2 \int W(x) f(x)\, \mathrm{d}x\{1 + \beta\gamma\rho^d / N\}, \tag{5.1}$$

on ignoring terms of lower order. Auestad and Tjøstheim (1990) have derived an expression similar to (5.1) using an argument similar to Akaike (1970) for the derivation of the now well-known final prediction error (FPE). This then shows the asymptotic equivalence of the CV and the FPE order determination criteria in the context of unknown autoregressive function F. However, note that, unlike the FPE approach, the CV approach does not invoke the assumption of independent copies used by Akaike (1970).

Cheng and Tong (1992, p. 434) have observed that if $k(0) = 0$, then under (A1)–(A14),

$$\mathrm{RSS} = \sigma_N^2\{1 + \beta\gamma\rho^d / N + \mathrm{o_p}(\rho^d / N)\}$$

on using Theorem 1 alone. Thus RSS with $k(0) = 0$ is equivalent to a special case of CV. An interesting question is thus: Does this observation provide a short cut to the proof of the result (equivalent to Theorem 2)

$$\mathrm{CV} = \sigma_N^2\{1 + \beta\gamma\rho^d / N + \mathrm{o_p}(\rho^d / N)\}$$

for *general* k subject to only (A1)–(A14), i.e. without restricting $k(0) = 0$? The obvious device of replacing k by k' with

$$k'(x) = \begin{cases} k(x), & x \neq 0, \\ 0, & x = 0, \end{cases}$$

would unfortunately require the relaxation of (A7). It would not seem difficult to relax (A7) to the extent that the inequality is only required to hold almost surely. However, the 'event $y = Y_s$' in (1.3)–(1.4) may have non-zero probability as $N \to \infty$.

Another related issue is the so-called predictive residual sum of squares (PRE). This may be obtained by replacing $\hat{f}$ by $\hat{f}_t$ and $\hat{F}$ by $\hat{F}_t$, namely

$$\hat{f}_t(y) = \frac{1}{(t-r+1)B^d} \sum_{s=r}^{t} K\left(\frac{y - Y_s}{B}\right)$$

and

$$\hat{F}_t(y) = \frac{1}{(t-r+1)B^d} \sum_{s=r}^{t} Z_s K\left(\frac{y - Y_s}{B}\right)\{\hat{f}_t(y)\}^{-1},$$

and

$$\mathrm{PRE}(d) = n^{-1} \sum_{t=r}^{N} \{Z_t - \hat{F}_t(Y_t)\}^2 W(Y_t).$$

Unfortunately we have not been able to obtain any theoretical results yet but would offer the obvious conjecture that its behaviour will be connected with the well-known Bayesian information criterion in conventional linear time series analysis.

Acknowledgements

BC thanks the Royal Society (UK) for financial support and Professor P.M. Robinson for his kindness and guidance during his visit to the London School of Economics. HT thanks the Science and Engineering Research Council for support, firstly for funds which enabled him to organise an international workshop in non-linear time series, held at Edinburgh in July 1989, from which much of the original stimulus for this paper was derived, and secondly for funds from their Complex Stochastic Systems Initiative. We thank the two referees for their very careful reading and constructive comments.

References

H. Akaike, Statistical prediction identification, Ann. Inst. Statist. Math. 22 (1970) 203–217.

Hong-Zhi An, Zhao-Guo Chen and E.J. Hannan, Autocorrelation, autoregression and autoregressive approximation, Ann. Statist. 10 (1982) 926–37.

B. Auestad and D. Tjøstheim, Identification of nonlinear series: first order characterization and order determination, Biometrika, 77 (1990) 669–688.

R. Bradley, Basic properties of strong mixing conditions, in E. Eberlein and M.S. Taqqu, eds., Dependence in Probability and Statistics (Birhauser, Boston, MA, 1986).

B. Cheng and H. Tong, On non-parametric order determination and chaos, J. Roy. Statist. Soc. Ser. B 54 (1992) 427–449.

B. Cheng and H. Tong, Nonparametric function estimation in noisy chaos, in: T. Subba Rao, ed., Birthday Volume of M.B. Priestly (Chapman and Hall, London), to appear.

M. Denker and G. Keller, On U-statistics and von Mises' statistics for weakly dependent processes, Z. Wahrsch. Verw. Gebiete 64 (1983) 505–522.

L. Gjörfi, W. Härdle, P. Sarda and P. Vieu, Nonparametric Curve Estimation From Time Series. Lecture Notes in Statist. (Springer, Berlin, 1990).
E.J. Hannan and L. Kavalieris, Regression, autoregression models, J. Time Ser. Anal. 7 (1986) 27–49.
W. Härdle, Applied Nonparametric Regression (Cambridge Univ. Press, Cambridge, 1990).
W. Härdle, P. Hall and J.S. Marron, How far are automatically chosen regression smoothing parameters from their optimum? J. Amer. Statist. Assoc. 83 (1988) 86–101.
W. Härdle and J.S. Marron, Optimal bandwidth selection in nonparametric regression function estimation, Ann. Statist., 13 (1985) 1461–1481.
A. Mokkadem, Sur un modéle autorégressif non linéaire, ergodicité et ergodicité géométrique, J. Time Ser. Anal., 8 (1987) 195–204.
A. Mokkadem, Mixing properties of ARMA processes, Stochastic Process. Appl. 29 (1988) 309–315.
E.A. Nadaraya, On nonparametric estimation of density function and regression, Theory Probab. Appl. 10 (1965) 186–190.
D.T. Pham, The mixing property of bilinear and generalized random coefficient autoregressive models, Stochastic Process. Appl. 23 (1986) 291–300.
D.T. Pham and L.T. Tran, Some mixing properties of time series models, Stochastic Process. Appl. 19 (1985) 297–303.
P.M. Robinson, Root-N-consistent semi-parametric regression, Econometrica, 56 (1988) 931–954.
M. Rosenblatt, Conditional probability density and regression estimators, Multivariate Anal. 2 (1969) 25–31.
G.G. Roussas, Nonparametric estimation in Markov processes, Ann. Inst. Statist. Math. 21 (1969) 73–87.
G.G. Roussas, Nonparametric estimation in mixing sequences of random variables, J. Statist. Plann. Inference 15 (1988) 135–149.
G.S. Watson, Smooth regression analysis. Sankhyā Ser. A 26 (1964) 359–372.